Biol Fertil Soils (2017) 53:479–484
DOI 10.1007/s00374-017-1191-3

EDITORIAL

Making big data smart—how to use metagenomics to understand soil quality

Gisle Vestergaard[1] · **Stefanie Schulz**[1] · **Anne Schöler**[1] · **Michael Schloter**[1,2]

Received: 12 June 2016 / Revised: 11 February 2017 / Accepted: 7 March 2017 / Published online: 16 March 2017

Introduction

Next-generation sequencing (NGS) has revolutionized the field of biology over the last decade. The Genomes OnLine Database (GOLD) that monitors sequencing projects worldwide has grown from just 1575 sequencing projects in 2005 to over 70,000 in 2015 (Reddy et al. 2015). This is partly caused by a rapid drop in the price of high-throughput sequencing (Hayden 2014), but also an increase of free user-friendly bioinformatical tools such as MG-RAST (Meyer et al. 2008), MEGAN (Huson et al. 2016) and user fora such as seqanswers.com, biostars.org etc.

This "brave new world" was introduced into soil sciences more than 10 years ago (Daniel 2005) and is becoming increasingly popular, as it is the only approach known, which allows a direct assessment of microbial community composition and function on various trophic levels. Today, according to the web of science, more than 900 papers have been published on soil metagenomes. In early times, sequencing depth was in the range of less than 1 Gbase and often resulted in the identification of only major functional traits and house keeping genes; today in recent publications up to 100 Gbases have been sequenced (Hultman et al. 2015), which allowed even a partly reconstruction of genomes of single microbes from the obtained reads. However, the interpretation of soil metagenomics data is still a challenge, given the often complex composition of the microbiomes, as well as their huge dynamics in time and space (Ebrahimi and Or 2016).

Previous papers have focused on specific aspects of metagenomic data generation or analysis, such as the impact of the DNA extraction methods and read annotation stringency on the apparent composition of a metagenome (Delmont et al. 2013), the importance of coverage estimation (Rodriguez-R and Konstantinidis 2014a) or the change from the current use of gene-centric snapshots towards genome-centric temporal studies (Prosser 2015). Major steps and recommendations for further reading are summarized in Table 1. In this paper, we discuss some basic guidelines for the experimental design of metagenomic surveys to characterize community composition and function of soil microbiomes, without losing the environmental context.

Sampling strategy

Soils are vertically and horizontally structured ecosystems, which are composed of a multitude of different microhabitats comprising diverse physical, chemical, and biological properties (Totsche et al. 2010). The degree of heterogeneity strongly depends on (i) the sampled compartment, e.g., the rhizosphere is less heterogeneous compared to bulk soil (Hinsinger et al. 2009), (ii) the soil texture, which strongly influences aggregate formation and also nucleic acid extraction efficiency, (iii) the above ground diversity and plant coverage, (iv) season, and (v) specific site characteristics like slope, shadowing, and groundwater table. (Petersen and Esbensen 2005). Taking this heterogeneity into account, the typically 500 mg to 10 g soil used for DNA extraction often do not reflect a single microsite, but a mixture of different compartments with differing chemical, physical, and biological properties, which often makes data interpretation quite challenging and only allows a correlative analysis of microbial data with abiotic soil

✉ Michael Schloter
schloter@helmholtz-muenchen.de

[1] Research Unit for Environmental Genomics, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[2] Chair for Soil Science, Technische Universität München, Munich, Germany

Table 1 Checklist for analysis of metagenomic datasets

NGS checklist		
Major step	Essential parts	Recommended references
Soil Sampling	-Determine soil type/texture, sampling date, pH, study specific parameters -Include negative control samples -Sample at least 3 replicates (each consisting of composite soil samples) -Store samples cold, also during sampling	Albertsen et al. 2015; Gilbert et al. 2014; Penton et al. 2016; Prosser 2010; Salter et al. 2014; Tatangelo et al. 2014; Totsche et al. 2010
Sequencing library preparation	-Check for inhibitory effects -Avoid multiple displacement amplification (MDA) -Use your controls - Shear DNA for shotgun sequencing	Salter et al. 2014; Yilmaz et al. 2010
Bioinformatic data analysis	-Remove contaminants -Remove adapters -Quality & length filter -Estimate coverage -If possible, use mock community (defined mixture of microbial cells) to validate your workflow -Upload raw sequencing data to public server	Bergkemper et al. 2016; Darzi et al. 2016; Del Fabbro et al. 2013; Menzel et al. 2016; Rodriguez-R and Konstantinidis 2014a; Sanchez-Flores et al. 2015; Schmieder and Edwards 2011; Schubert et al. 2016; Wood and Salzberg 2014

properties, but does not increase our mechanistic understanding of how soil ecosystems work.

Although the effect of spatial heterogeneity can be reduced by increasing the amount of soil used for DNA extraction (Penton et al. 2016), resulting in the above-mentioned problems related to data interpretation, one soil sample taken from a given site is by far not representative and like in all other biological experiments true replicates need to be analyzed (Prosser 2010). In the best case, the underlying sampling design should include a geostatistical setting to better characterize the sampling site. In any case, a minimum of three replicates per treatment is needed to perform proper statistic testing. To identify the optimal number of replicates, a statistical power analysis can be performed, which reduces the chance of type II errors (Klironomos et al. 1999). However, often the result of such a power analysis might not meet the financial and computational frames of a sequencing project. Thus, to limit the influence of soil heterogeneity, a representative sampling strategy is often used, which includes a pooling of sub-samples (Fig. 1). Mainly for agricultural soils, besides horizontal distribution patterns, also issues related to different soil layers are of interest. Thus, the sampling strategy should be chosen based on the soil stratification to avoid the mixing of different soil horizons. Besides spatial issues also temporal issues need to be taken into account, as soil microbiomes change strongly in response to land management, like fertilization or tillage, plant development stage, climate, and season (Ollivier et al. 2011). Thus, one sampling time point does not help to understand the complexity of microbiomes at a given site but is often a snapshot, which is strongly influenced by the recent local conditions. Taking these strong dynamics of soil microbiomes in time and space (Kuzyakov and Blagodatskaya 2015) into account, the sampling strategy should be driven by a clear research hypothesis.

As abiotic soil parameters are a major driver of soil microbiomes, besides the factors of interest, a minimum dataset is required, which needs to be analyzed and implemented independently from the research questions. Besides exact GPS coordinates and climatic conditions at the period of sampling, such metadata should include the soil type, soil texture, soil pH, stable pools of soil organic matter like total organic C and N, and labile pools of C, N, P, and S. If agricultural sites are studied, management-related properties like fertilization regimes, tillage, cropping sequence, plant protection measures, and plant biomass should be given. For unmanaged sites at least above ground diversity should be characterized.

Sample processing and downstream analysis

Soils should be stored after sampling at a suitable temperature, which is below 4 °C for short-term storage in the field and −20 °C for long-term storage (Lauber et al. 2010; Tatangelo et al. 2014). Compared to amplicon-based sequencing, the direct DNA sequencing (metagenomics) requires higher amounts of high-quality DNA, which in turn also depends on the kit used for library preparation (500 pg–1 μg). Thus, there is often a need to adapt the used DNA extraction protocol to fulfill these requirements. The use of multiple displacement amplification should be avoided taking the significant bias introduced into account (Yilmaz et al. 2010). Since DNA extraction protocols vary in efficiency depending on the nature of the samples and in removing various inhibitors we recommend testing the workflow on a few non-essential samples first (Frostegård et al. 1999). After a DNA extraction method has been selected, it should be used consistently given

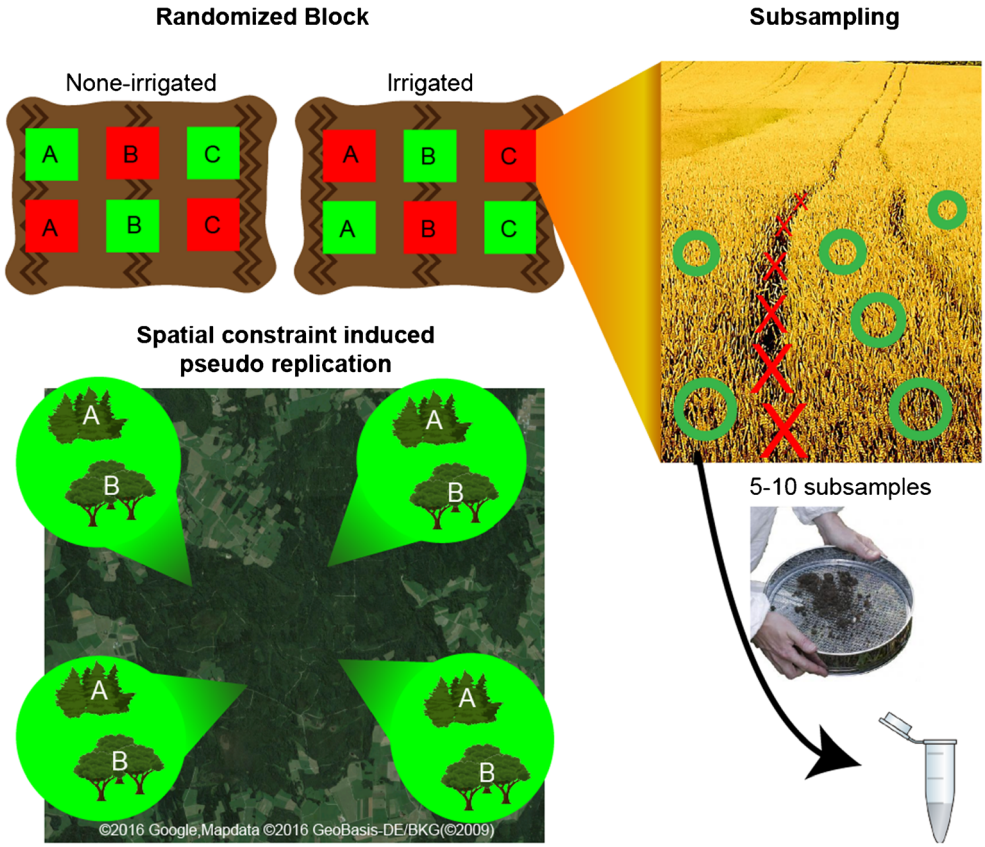

Fig. 1 Soil Sampling. "Randomized Block" and "Spatial constraint induced pseudo replication" are two common ways of soil sampling compromising practicality and reproducibility. Field studies are often divided into blocks for pragmatic reasons (plowing, irrigation etc.) while other variables here illustrated by *colors green* and *red* are more easily distributed randomly to increase reproducibility. *A*, *B*, and *C* indicate replicates. Spatial constraints are often imposed since many biological systems are not uniformly distributed such as pine trees in a wild forest. Given microenvironments etc., soil samples are optimally composed of several subsamples which are not anomalous, such as in the shown examples, wheel tracks

the inherent bias introduced throughout the whole project. Finally, depending on the aim of the study, one might also consider employing methods that separate extracellular DNA from intracellular DNA (Pietramellara et al. 2009) which allow a discrimination between alive and dead microbes. As recommended by the Earth Microbiome Project (Gilbert et al. 2014) and due to the impact of downstream procedures like DNA extraction or library preparation on detected microbial communities (Albertsen et al. 2015), it is essential to include negative controls, e.g., negative DNA extractions (Salter et al. 2014), mainly if low amounts of DNA (<5 ng) are used for sequencing.

Rapid advances in sequencing technology, which each have their specific challenges, make it impossible to provide universal guidelines. With 454 pyrosequencing being outdated and long read technologies such as Oxford Nanopore Technologies and PacBio® yet not frequently used for metagenomics, here, we focus on Illumina-based technologies, which are currently the de facto standard in metagenomics (Sanchez-Flores et al. 2015).

The needed quality of reads obtained by sequencing is highly dependent on questions asked, but nevertheless quality filtering of the sequences is essential and should be adjusted specifically for the dataset at hand to optimize the trade-off between read-loss and final quality of the dataset (Del Fabbro et al. 2013). Key quality controls should include the following steps: removal of sequencing adapters, quality and length filtering, and removal of possible contaminants such as PhiX and/or host DNA. A good combination is adapter removal for the removal of adapters, quality/length trimming, and merging of paired sequences (Schubert et al. 2016), followed by Deconseq for the removal of contaminants (Schmieder and Edwards 2011). Lack of proper contaminant removal is especially critical with Illumina sequencing as apparent from the large scale contamination of microbial isolate genomes with Illumina PhiX control DNA (Mukherjee et al. 2015).

The sequencing depth for a sound bioinformatic analysis strongly depends on the aims of the project. If binning is planned to assemble larger contigs from the obtained reads, sequencing depth of up to 100 Gbases per sample are needed (Hultman et al. 2015), for a pure comparison of single reads, for example, to reconstruct major nutrient cycles in a given soil much lower sequencing depth (5–10 Gbases) are required (Bergkemper et al. 2016). While highly recommended, estimating the obtained sequencing depth or coverage of a metagenome is challenging compared to, e.g., 16S rRNA-based amplicon sequencing. Using16S rRNA-based amplicon sequencing, we can assume that public databases allow us to identify the vast majority of reads, while comparing metagenomic datasets to public databases such as the NCBI non-redundant protein database or functional assignment databases such as KEGG (Kanehisa et al. 2016), SEED

(Overbeek et al. 2005) or COG (Tatusov et al. 2000) would only identify a part of the reads and have a bias towards model and/or medically relevant organisms. Therefore, rarefaction analysis makes sense with 16S rRNA amplicons to assess species richness and sample coverage, while rarefaction of metagenomics datasets to assess metagenomic complexity and sample coverage would overestimate coverage, which is not even consistent across different samples. Thus, for more accurate coverage estimations of metagenomics data, database-independent approaches are needed. Nonpareil (Rodriguez-R and Konstantinidis 2014b), which examines the degree of overlap among individual sequences to assess if a sufficient coverage has been achieved, is a good alternative to overcome the above-mentioned problems (Rodriguez-R and Konstantinidis 2014a).

Assembling contigs from reads can significantly increase the quality of annotation, especially when working with the shorter reads provided by the HiSeq platforms. Assembly programs such as IDBA-UD (Peng et al. 2012) and MegaHit (Li et al. 2016) provide well-established pipelines which are also well accepted in literature. While general functional annotation databases such as the aforementioned are useful for descriptive studies and to obtain a broad overview of the data, they are often based on eukaryotic or model organisms leading to suboptimal functional assignments (Darzi et al. 2016). Thus, more targeted approaches might be very useful, such as the FOAM database (Prestat et al. 2014), which was developed specifically to screen environmental metagenomic data and is an improvement for any soil-related study. For studies of particular genes of interest, even more, focused approaches and specialized databases are needed depending on the research question. Depending on the availability of such specialized databases, one should either use or create custom databases to compare the metagenomics sequences to and/or employ hidden Markov models to detect conserved domains in the metagenomics sequences. Combining an initial metagenomic screen with subsequent amplicon sequencing can in some cases further increase sensitivity albeit often at a cost of limiting diversity (Bergkemper et al. 2016). For assembly-free taxonomic classification, several solutions are recommendable such as Kraken and Kaiju (Wood and Salzberg 2014; Menzel et al. 2016). In any case, the used bioinformatics pipeline must be well described as so far no "gold standard" for data analysis is available. The first data provided by the CAMI initiative (Critical Assessment of Metagenome Interpretation) has proven significant differences in the outcome of read analysis depending on the used software. In this respect, there is a need that sequences are deposited in public databases in their raw forms, as even data trimming introduces biases depending on the used method.

Outlook

Despite the ever growing sequence databases, most metagenomic reads cannot be assigned to a function, limiting both our ability to test hypotheses, but also the value of metagenomic datasets as a tool for novel discoveries. Besides developing targeted approaches for the isolation of microorganisms from soil, which allows a classical taxonomic assignment of genotypic and phenotypic traits, novel approaches integrating metagenomic datasets with other types of data such as metabolomics and abiotic factors are starting to yield much greater insight into the workings of the microbiome (Feng et al. 2016).

As the analysis of DNA provides a potential for the expression of certain genes only, there has been a great interest in applying a comparable pipeline like described above for the analysis of metatranscriptomes from soil (Baldrian et al. 2012). In principle, the same approach can be also used for the analysis of extracted RNA from soil after reverse transcription. Due to the high stability of rRNA compared to mRNA, depletion techniques are needed to reduce the amount of rRNA. Furthermore, the issue of spatial and temporal heterogeneity is more pronounced when analyzing RNA, as the stability of mRNA in cells is often in the order of minutes to hours, thus one sampling may reflect only a snapshot depending the actual environmental conditions.

Moreover, the development of long-read sequencing technologies opens a new field of application, which has the potential to provide additional information about operon structures from samples with low diversity or samples where a specific target was enriched beforehand. Such approaches will help us in the future to improve our understanding on mechanisms how gene expression is regulated opening a new field in soil microbial ecology addressing issues of "metaregulation." Such studies could help us to improve our understanding for example on the molecular mechanisms of major ecosystem services provided by soils like plant growth promotion or carbon sequestration.

Acknowledgements Gisle Vestergaard is supported by a Humboldt Research Fellowship for postdoctoral researchers.

References

Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH (2015) Back to basics—the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One. doi:10.1371/journal.pone.0132783

Baldrian P, Kolarik M, Stursova M, Kopecky J, Valaskova V, Vetrovsky T, Zifcakova L, Snajdr J, Ridl J, Vlcek C, Voriskova J (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J 6:248–258. doi:10.1038/ismej.2011.95

Bergkemper F, Kublik S, Lang F, Krüger J, Vestergaard G, Schloter M, Schulz S (2016) Novel oligonucleotide primers reveal a high diversity of microbes which drive phosphorous turnover in soil. J Microbiol Methods 125:91–97. doi:10.1016/j.mimet.2016.04.011

Daniel R (2005) The metagenomics of soil. Nat Rev Microbiol 3:470–478. doi:10.1038/nrmicro1160

Darzi Y, Falony G, Vieira-Silva S, Raes J (2016) Towards biome-specific analysis of meta-omics data. ISME J 10:1025–1028. doi:10.1038/ismej.2015.188

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. doi:10.1371/journal.pone.0085024

Delmont TO, Simonet P, Vogel TM (2013) Mastering methodological pitfalls for surviving the metagenomic jungle. BioEssays: news Rev Mol Cell Dev Biol 35:744–754. doi:10.1002/bies.201200155

Ebrahimi A, Or D (2016) Microbial community dynamics in soil aggregates shape biogeochemical gas fluxes from soil profiles - upscaling an aggregate biophysical model. Glob Chang Biol 3141–3156–3141–3156. doi: 10.1111/gcb.13345

Feng Q, Liu Z, Zhong S, Li R, Xia H, Jie Z, Wen B, Chen X, Yan W, Fan Y, Guo Z, Meng N, Chen J, Yu X, Zhang Z, Kristiansen K, Wang J, Xu X, He K, Li G (2016) Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease. Sci reports. doi:10.1038/srep22525

Frostegård A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, Jeannin P, Nesme X, Simonet P (1999) Quantification of bias related to the extraction of DNA directly from soils. Appl Environ Microbiol 65:5409–5420

Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. BMC Biol. doi:10.1186/s12915-014-0069-1

Hayden EC (2014) Technology: the $1,000 genome. Nature 507:294–295. doi:10.1038/507294a

Hinsinger P, Bengough AG, Vetterlein D, Young IM (2009) Rhizosphere: biophysics, biogeochemistry and ecological relevance. Plant Soil 321:117–152. doi:10.1007/s11104-008-9885-9

Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB, VerBerkmoes NC, Lee LH, Mavrommatis K, Jansson JK (2015) Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature 521:208–212. doi:10.1038/nature14238

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R (2016) MEGAN community edition—interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. doi:10.1371/journal.pcbi.1004957

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462. doi:10.1093/nar/gkv1070

Klironomos JN, Rillig MC, Allen MF (1999) Designing belowground field experiments with the help of semi-variance and power analyses. Appl Soil Ecol 12:227–238. doi:10.1016/S0929-1393(99)00014-1

Kuzyakov Y, Blagodatskaya E (2015) Microbial hotspots and hot moments in soil: concept & review. Soil Biol Biochem 83:184–199. doi:10.1016/j.soilbio.2015.01.025

Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307:80–86. doi:10.1111/j.1574-6968.2010.01965.x

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W (2016) MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11. doi:10.1016/j.ymeth.2016.02.020

Menzel P, Ng KL, Krogh A (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. doi:10.1038/ncomms11257

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinforma. doi:10.1186/1471-2105-9-386

Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A (2015) Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci. doi:10.1186/1944-3277-10-18

Ollivier J, Töwe S, Bannert A, Hai B, Kastl E-M, Meyer A, Su MX, Kleineidam K, Schloter M (2011) Nitrogen turnover in soil and global change. FEMS Microbiol Ecol 78:3–16. doi:10.1111/j.1574-6941.2011.01165.x

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691–5702. doi:10.1093/nar/gki866

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinforma 28:1420–1428. doi:10.1093/bioinformatics/bts174

Penton CR, Gupta VVSR, Yu J, Tiedje JM (2016) Size matters: assessing optimum soil sample size for fungal and bacterial community structure analyses using high throughput sequencing of rRNA Gene amplicons. Front Microbiol. doi:10.3389/fmicb.2016.00824

Petersen L, Esbensen KH (2005) Representative process sampling for reliable data analysis—a tutorial. J Chemom 19:625–647. doi:10.1002/cem.968

Pietramellara G, Guerri G, Ascher J, Borgogni F, Nannipieri P, Ceccherini MT (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. Biol Fertil Soils 45:219–235. doi:10.1007/s00374-008-0345-8

Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J, Mackelprang R, Myrold DD, Jumpponen A, Tringe SG, Holman E, Mavromatis K, Jansson JK (2014) FOAM (functional ontology assignments for metagenomes): a hidden Markov model (HMM) database with environmental focus. Nucleic Acids Res 42:e145–e145. doi:10.1093/nar/gku702

Prosser JI (2010) Replicate or lie. Environ Microbiol 12:1806–1810. doi:10.1111/j.1462-2920.2010.02201.x

Prosser JI (2015) Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology. Nat Rev Microbiol 13:439–446. doi:10.1038/nrmicro3468

Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC (2015) The genomes OnLine database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res 43:D1099–D1106. doi:10.1093/nar/gku950

Rodriguez-R LM, Konstantinidis KT (2014a) Estimating coverage in metagenomic data sets and why it matters. ISME J 8:2349–2351. doi:10.1038/ismej.2014.76

Rodriguez-R LM, Konstantinidis KT (2014b) Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinforma 30:629–635. doi:10.1093/bioinformatics/btt584

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014)

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. doi:10.1186/s12915-014-0087-z

Sanchez-Flores A, Vera-Ponce de León A, Escobar-Zepeda A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet. doi:10.3389/fgene.2015.00348

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. doi:10.1371/journal.pone.0017288

Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res notes. doi:10.1186/s13104-016-1900-2

Tatangelo V, Franzetti A, Gandolfi I, Bestetti G, Ambrosini R (2014) Effect of preservation method on the assessment of bacterial community structure in soil and water samples. FEMS Microbiol Lett 356:32–38. doi:10.1111/1574-6968.12475

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. doi:10.1093/nar/28.1.33

Totsche KU, Rennert T, Gerzabek MH, Kögel-Knabner I, Smalla K, Spiteller M, Vogel H (2010) Biogeochemical interfaces in soil: the interdisciplinary challenge for soil science. J Plant Nutr Soil Sci 173:88–99. doi:10.1002/jpln.200900105

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. doi:10.1186/gb-2014-15-3-r46

Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods 7:943–944. doi:10.1038/nmeth1210-943